Using InfiniBand for a scalable compute infrastructure

Technology brief, 4th edition

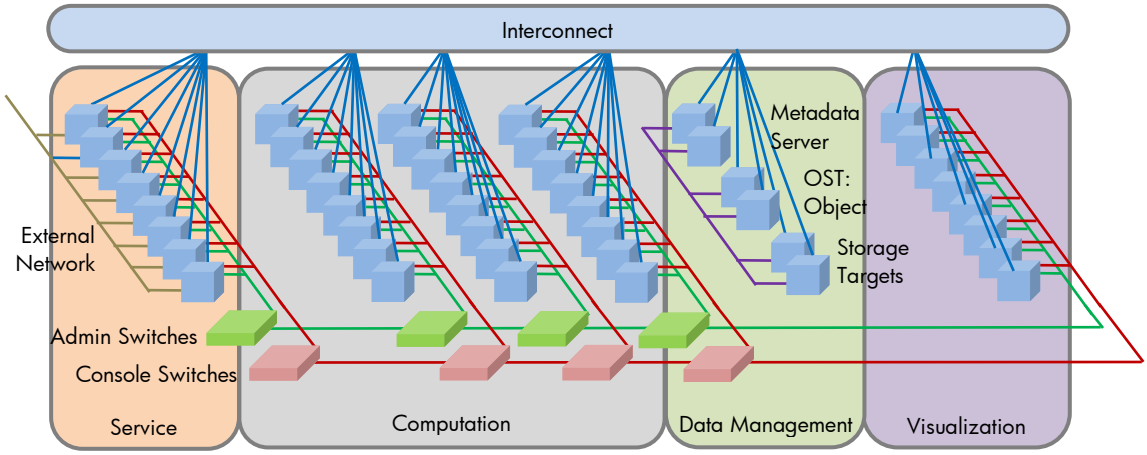
Introduction

Increasingly, IT organizations are deploying scale-out computing using clusters for these applications:

- High performance computing (HPC)
- Enterprise applications
- Financial services
- Scale-out databases

Scale-out computing uses interconnect technology to combine a large number of stand-alone server nodes into an integrated and centrally managed system (Figure 1). A cluster infrastructure works best when built with an interconnect technology that scales easily, reliably, and economically with system expansion.

Figure 1. System architecture for a sample scale-out HPC cluster



You can now choose from three interconnect technologies to build a cluster:

- Gigabit Ethernet (GbE)
- 10 GbE
- InfiniBand (IB)

Ethernet can be cost-effective for scale-out systems that run applications with little inter-node communication. 10 GbE meets higher bandwidth requirements than GbE can provide. InfiniBand is ideal for scale-out environments where applications have extensive inter-node communication and require very low latency and high bandwidth across the entire fabric.

This technology brief describes InfiniBand as an interconnect technology used in cluster computing. It describes multiple InfiniBand topologies and multiple configurations for building an HPC cluster. It also explains how InfiniBand achieves high performance in real world applications.

InfiniBand technology

InfiniBand is an industry-standard, channel-based architecture with an application-centric view to provide an easy-to-use messaging service. InfiniBand uses techniques such as stack bypass using remote direct memory access (RDMA) to let applications directly communicate with each other across the wire. This communication results in high-speed, low-latency connections for scale-out computing.

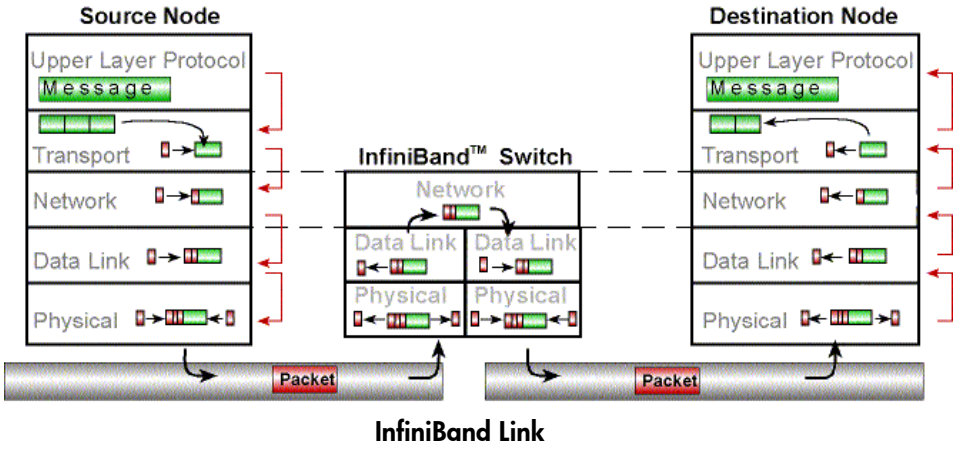
InfiniBand architecture

Like Ethernet, InfiniBand uses a multi-layer processing stack to transfer data between nodes (Figure 2). However, InfiniBand architecture includes OS-bypass functions such as communication processing duties and RDMA operations as core capabilities. InfiniBand offers greater adaptability through a variety of services and protocols.

The upper layer protocols work closest to the operating system and application. They define the services and affect how much software overhead data transfer will require. The InfiniBand transport layer is responsible for communication between applications. The transport layer splits the messages into data payloads. It encapsulates each data payload and an identifier of the destination node into one or more packets. Packets can contain data payloads of up to 4 kilobytes.

The network layer selects a route to the destination node and attaches the route information to the packets. The data link layer attaches a local identifier (LID) to the packet for communication at the subnet level. The physical layer transforms the packet into an electromagnetic signal based on the type of network media—copper or fiber.

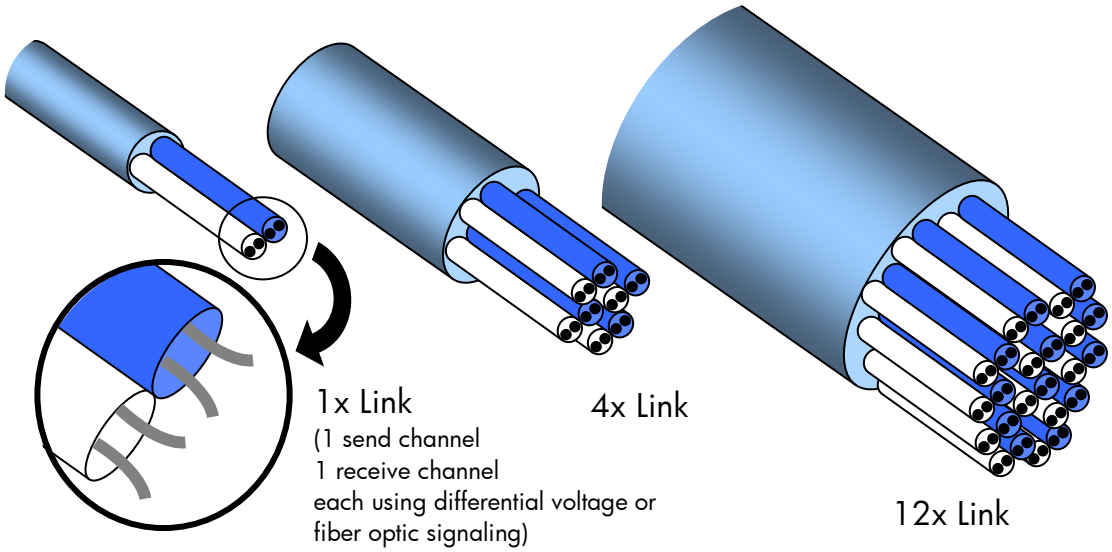
Figure 2. InfiniBand functional architecture



InfiniBand fabrics use high-speed, bi-directional serial interconnects between devices. Interconnect bandwidth and distance limits are determined by the type of cabling and connections used. The bi-directional links contain dedicated send and receive lanes for full duplex operation.

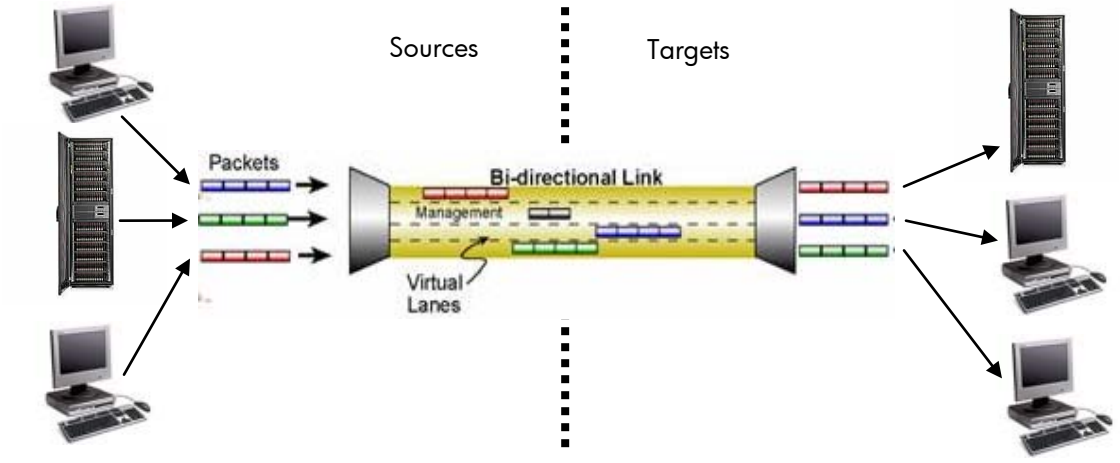
InfiniBand interconnect types include 1x, 4x, or 12x wide full-duplex links (Figure 3). The most popular configuration is 4x. It provides a theoretical full-duplex quad data rate (QDR) bandwidth of 80 (2 x 40) gigabits per second.

Figure 3. InfiniBand link types



The host channel adapter (HCA) can divide (multiplex) each link into a set of virtual lanes, similar to highway lanes (Figure 4). Using virtual lanes, an InfiniBand link can share bandwidth between various sources and targets simultaneously. For example, if the HCA divides a 10 Gb/s link into five virtual lanes, each lane has a bandwidth of 2 Gb/s. The InfiniBand architecture defines a virtual lane-mapping algorithm to ensure inter-operability between end nodes that support different numbers of virtual lanes.

Figure 4. InfiniBand virtual lane operation



Overhead in data transmission limits the maximum data bandwidth per link to a peak of 80 percent of the signal rate. However, InfiniBand's switched fabric design lets bandwidth scale as you add links and nodes.

InfiniBand Quality of Service functions

All traffic in a clustered environment competes for fabric resources such as physical links and queues. To minimize interference and congestion, InfiniBand implements Quality of Service (QoS) functions on each virtual lane.

Flow control

Ethernet uses flow control as a time-based function. To avoid input overload, the receiving NIC tells the sending NIC to stop sending for a time. The protocol does not guarantee packet delivery; the NIC must re-send dropped packets.

InfiniBand achieves flow control through a point-to-point, credit-based scheme reflected within each packet. The HCA of each target puts available buffer resource information into a credit counter. The source HCA reads available credits reported by the target HCA. The source HCA does not send a packet until the target HCA indicates that space is available for it. There are no dropped packets.

Congestion management

InfiniBand congestion management identifies congestion hot spots. It applies advanced routing techniques such as traffic-aware algorithms to improve throughput.

Service differentiation

InfiniBand organizes data by classifying traffic into service levels. The HCA assigns each packet a service level. It also maps each service level to a virtual lane. The HCA assigns a priority to each virtual lane to give high-priority traffic the advantage, but not domination, over other traffic types. A weighted arbitration scheme allows virtual lanes of the same priority an even chance to transfer traffic.

InfiniBand performance

HPC interconnect performance is commonly measured by latency and bandwidth.

Latency

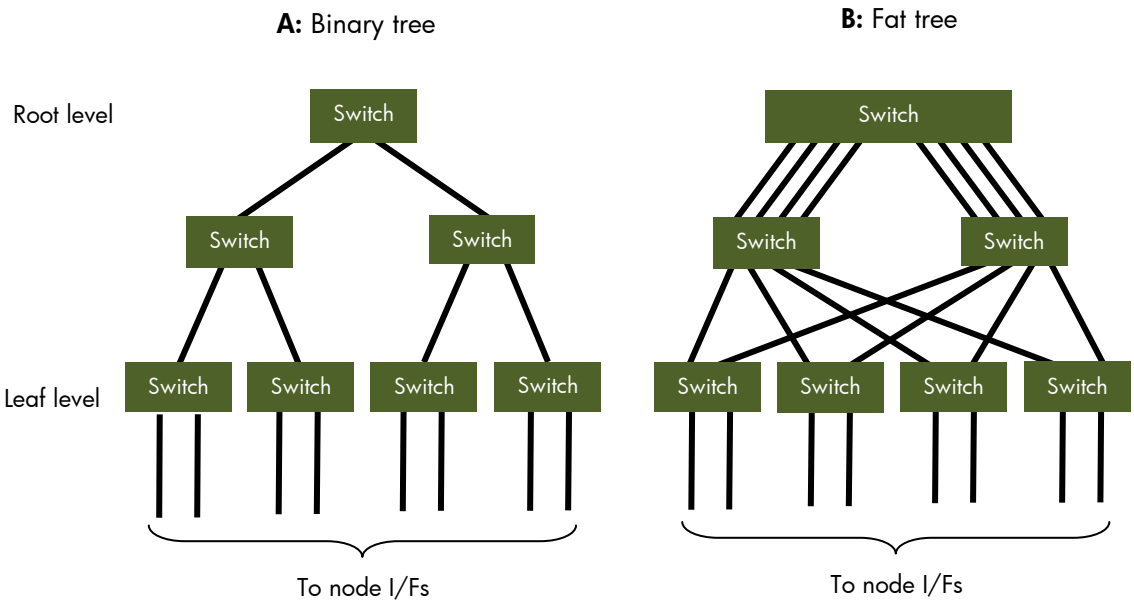
Latency is the time required to transfer a packet from a process on the source node to a process on the target node.

Low latency is critical for high application performance because the data in a packet is unavailable to the application while in route. Latency for a single core processing a packet over 10 GbE can be five to six times greater than for an InfiniBand transaction. Many HPC applications use multiple cores, which increases the importance of latency. In tests using Intel® MPI benchmarks, even minimal core scaling (up to eight cores) per node can increase latency as much as 60 percent on a10 GbE system. The effect on an equivalent InfiniBand system is negligible for the same test.

Bandwidth

A network's bandwidth is a measure of its data throughput. Data rate generally determines network bandwidth, but network topology can affect bandwidth significantly. In a basic binary tree topology (Figure 5A), each progression from the leaf to root level reduces the aggregate bandwidth by half. Ethernet networks often have this topology. It creates an over-subscribed hierarchy resulting in severe congestion. HPC clusters typically use a fat-tree topology (Figure 5B) with parallel paths between levels to maintain a constant bisectional bandwidth.

Figure 5. InfiniBand cluster topologies

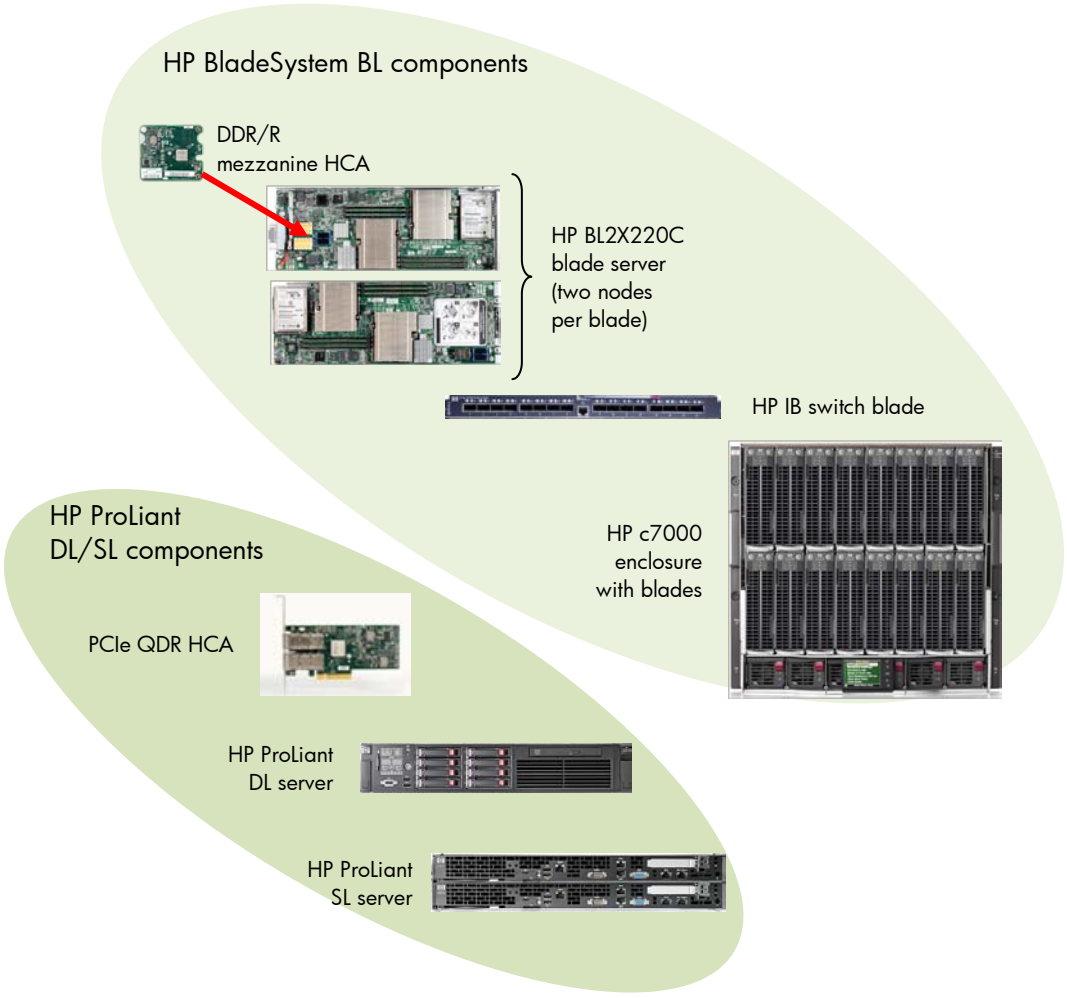


InfiniBand provides native support for fat-tree and other mesh topologies, allowing simultaneous connections across multiple links. The fabric scales as you connect more nodes and additional links.

Scale-out clusters built on InfiniBand and HP technology

Scale-out cluster computing has become the common architecture for HPC, and broader markets are adopting it. The trend is toward using space- and power-efficient systems for scale-out solutions. Figure 6 shows the key components for building a scalable HPC solution.

Figure 6. Solutions for scale-out HPC clusters

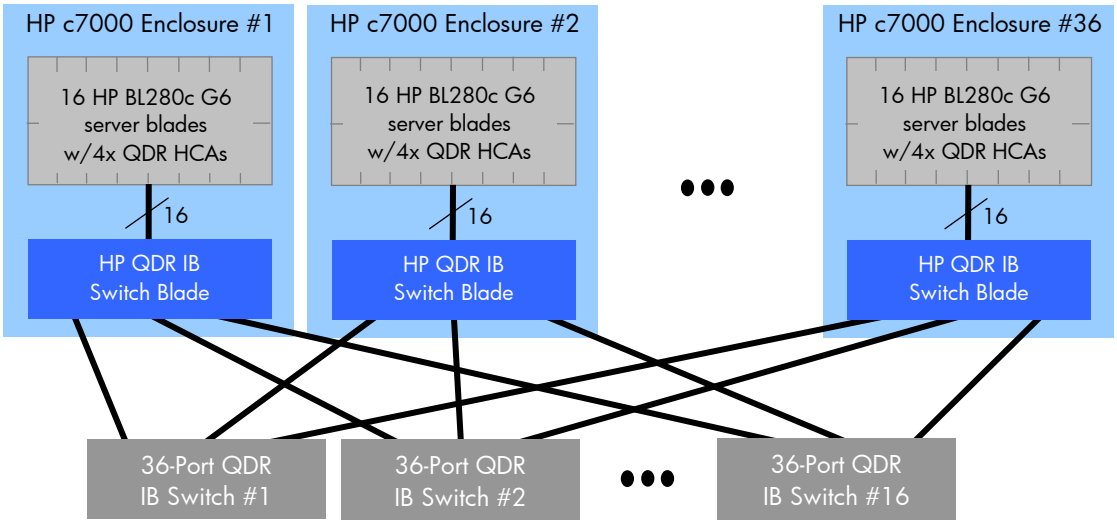


The following examples describe possible solutions for a 576-node HPC configuration.

HPC configuration with HP BladeSystem solutions

Figure 7 shows a full-bandwidth, fat-tree configuration of HP BladeSystem c-Class components providing 576 nodes in a cluster. Each c7000 enclosure includes an HP 4x QDR InfiniBand Switch Blade, with 16 downlinks for server blade connection and 16 QSFP uplinks for fabric connectivity. Sixteen 36-port QDR InfiniBand switches provide spine-level fabric connectivity.

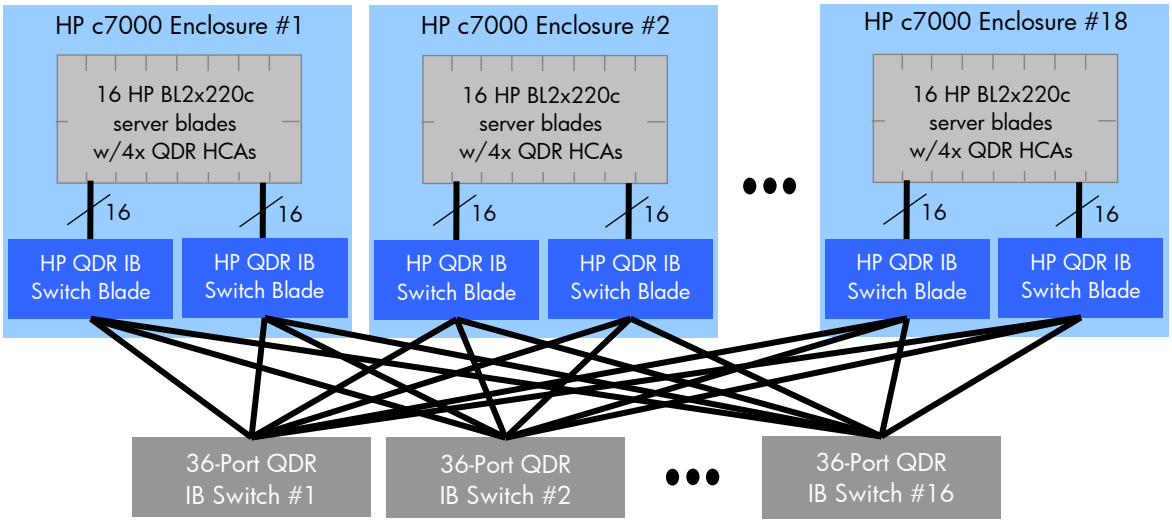
Figure 7. HP BladeSystem c-Class 576-node cluster configuration using BL280c blades



Total nodes	576 (1 per blade)
Racks required for servers	Nine 42U (assumes four c7000 enclosures per rack)
Interconnect	1:1 full bandwidth (non-blocking), 3 switch hops maximum, fabric redundancy

To meet extreme density goals, the half-height HP BL2x220c server blade includes two server nodes. Each node can support two quad-core Intel® Xeon® 5400-series processors and a slot for a mezzanine board. That equals up to 32 nodes (256 cores) per c7000 enclosure. Each c7000 enclosure contains two HP 4x QDR InfiniBand Switch Blades. Figure 8 shows how using the dual-node BL2x220c blade lets you deploy 576 nodes in half as much rack space.

Figure 8. HP BladeSystem c-Class 576-node cluster configuration using BL2x220c blades



Total nodes	576 (2 per blade)
Racks required for servers	Four 42U, one 36U (assumes four c7000 enclosures per rack)
Interconnect	1:1 full bandwidth (non-blocking), 3 switch hops maximum, fabric redundancy

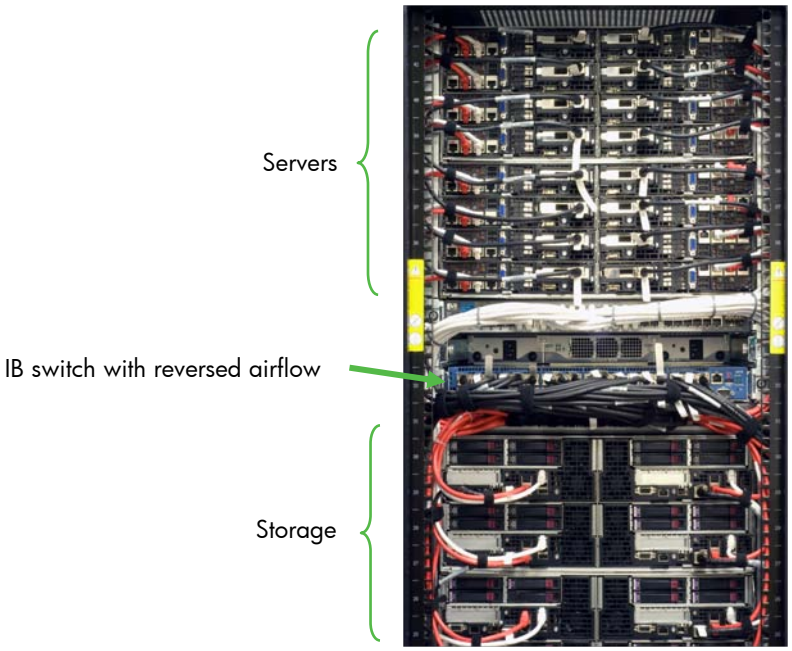
HPC-optimized HP Cluster Platforms

Deploying the HP Unified Cluster Portfolio (UCP) can give you the performance and flexibility of a custom solution with the simplicity and value of a factory-built product. The HP UCP is a modular package of hardware, software, and services.

You can configure HP Cluster Platforms with HP ProLiant BL, DL, or SL systems. You have a choice of packaging styles, processor types, and density levels. You can choose InfiniBand solutions from Mellanox, Voltaire, or QLogic to meet your latency and bandwidth requirements.

We have optimized HP Cluster Platforms for both performance and serviceability. Figure 9 shows a partial rear view of an HP Cluster Platform rack. InfiniBand switches with reversed (back-to-front) airflow are mounted in the rear of the rack. Note how this configuration allows well-organized cabling for ease of serviceability and non-restrictive airflow.

Figure 9. Partial rear view of an HP Cluster Platform rack



Conclusion

You should base your decision to use Ethernet or InfiniBand on performance and cost requirements. We are committed to supporting both InfiniBand and Ethernet infrastructures. We want to help you choose the most cost-effective fabric solution for your environment.

InfiniBand is the best choice for HPC clusters requiring scalability from hundreds to thousands of nodes. While you can apply zero-copy (RDMA) protocols to TCP/IP networks such as Ethernet, RDMA is a core capability of InfiniBand architecture. Flow control and congestion avoidance are native to InfiniBand.

InfiniBand also includes support for fat-tree and other mesh topologies that allow simultaneous connections across multiple links. This lets the InfiniBand fabric scale as you connect more nodes and links.

Parallel computing applications that involve a high degree of message passing between nodes benefit significantly from InfiniBand. Data centers worldwide have deployed DDR for years and are quickly adopting QDR. HP BladeSystem c-Class clusters and similar rack-mounted clusters support IB DDR and QDR HCAs and switches.

For more information

Resource description	Web address
HP products	www.hp.com
HPC/IB/cluster products	www.hp.com/go/hptc
HP InfiniBand products	http://h18004.www1.hp.com/products/servers/networking/index-ib.html
InfiniBand Trade Organization	http://www.infinibandta.org
Open Fabrics Alliance	http://www.openib.org/
RDMA Consortium	http://www.rdmaconsortium.org.
Technology brief discussing iWARP RDMA	http://h20000.www2.hp.com/bc/docs/support/SupportManual/c00589475/c00589475.pdf
HP BladeSystem	http://h18004.www1.hp.com/products/blades/components/c-class-tech-function.html

Call to action

Send comments about this paper to TechCom@HP.com

TC101215TB December 2010